Search Engines

II. How Software Agents and Search Engines Work

There are at least three elements to search engines that I think are important: information discovery & the database, the user search, and the presentation and ranking of results.

Discovery and Database

A search engine finds information for its database by accepting listings sent in by authors wanting exposure, or by getting the information from their "Web crawlers," "spiders," or "robots," programs that roam the Internet storing links to and information about each page they visit. Web crawler programs are a subset of "software agents," programs with an unusual degree of autonomy which perform tasks for the user. How do these really work? Do they go across the net by IP number one by one? Do they store all or most of everything on the Web?

According to The WWW Robot Page, these agents normally start with a historical list of links, such as server lists, and lists of the most popular or best sites, and follow the links on these pages to find more links to add to the database. This makes most engines, without a doubt, biased toward more popular sites. A Web crawler could send back just the title and URL of each page it visits, or just parse some HTML tags, or it could send back the entire text of each page. Alta Vista is clearly hell-bent on indexing anything and everything, with over 30 million pages indexed (7/96). Excite actually claims more pages. OpenText, on the other hand, indexes the full text of less than a million pages (5/96), but stores many more URLs. Inktomi has implemented HotBot as a distributed computing solution, which they claim can grow with the Web and index it in entirety no matter how many users or how many pages are on the Web. By the way, in case you are worrying about

software agents taking over the world, or your Web site, look over the Robot Attack Page. Normally, "good" robots can be excluded by a bit of Exclusion Standard code on your site.

It seems unfair, but developers aren't rewarded much by location services for sending in the URLs of their pages for indexing. The typical time from sending your URL in to getting it into the database seems to be 6-8 weeks. Not only that, but a submission for one of my sites expired very rapidly, no longer appearing in searches after a month or two, apparently because I didn't update it often enough. Most search engines check their databases to see if URLs still exist and to see if they are recently updated.

User Search

What can the user do besides typing a few relevant words into the search form? Can they specify that words must be in the title of a page? What about specifying that words must be in an URL, or perhaps in a special HTML tag? Can they use all logical operators between words like AND, OR, and NOT?

Most engines allow you to type in a few words, and then search for occurrences of these words in their data base. Each one has their own way of deciding what to do about approximate spellings, plural variations, and truncation. If you just type words into the "basic search" interface you get from the search engine's main page, you also can get different logical expressions binding the different words together. Excite! actually uses a kind of "fuzzy" logic, searching for the AND of multiple words as well as the OR of the words. Most engines have separate advanced search forms where you can be more specific, and form complex Boolean searches (every one mentioned in this article except Hotbot). Some search tools parse HTML tags, allowing you to look for things specifically as links, or as a title or URL without consideration of the text on the page.

Query Syntax Checklist

How does your engine handle:

Truncation, Pluralization & Capitalization:
Macintosh, Mac, Macintoshes, Macs, macintosh, macintoshes, mac, macs, could all yield different results. Most engines interpret lower case as unspecified, but upper case will match only upper case, but there are exceptions. There is no standard at all for truncation, and worse yet, it is probably different in general and advanced search mode for every engine.

Multiple Words
does the engine

By searching only in titles, one can eliminate pages with only brief mentions of a concept, and only retrieve pages that really focus on your concept.

By searching links, one can determine how many and which pages point at your site. Understanding what each page does with the non-standard pluralization, truncation, etc. can be quite important in how successful your searches will be. For example, if you search for "bikes" you won't get "bicycle," "bicycles," or "bike." In this case, I would use a search engine that allowed "truncation," that is, one that allowed the search word "bike" to match "bikes" as well, and I would search for "bicycyle OR bike OR cycle" ("bicycle* OR bike* OR cycle*" in Alta Vista).

logically AND them or OR them ?

Phrases

Typically one puts quotes around a phrase so that each word in the phrase is not searched for separately.

. . . Check with your engine's help file before starting a search.

Presentation & Ranking

With databases that can keep the entire Web at the fingertips of the search engines, there will always be relevant pages, but how do you get rid of the less relevant and emphasize the more relevant?

Most engines find more sites from a typical search query than you could ever wade through. Search engines give each document they find some measure of the quality of the match to your search query, a relevance score. Relevance scores reflect the number of times a search term appears, if it appears in the title, if it appears at the beginning of the document, and if all the search terms are near each other; some details are given in engine help pages. Some engines allow the user to control the relevance score by giving different weights to each search word. One thing that all engines do, however, is to use alphabetical order at some point in their display algorithm. If relevance scores are not very different for various matches, then you end up with this sorry default. Zeb's [Whatever] page will never fare very well in this case, regardless of the quality of its content. For most uses, a good summary is more useful than a ranking. The summary is usually composed of the title of a document and some text from the beginning of the document, but can include an author-specified summary given in a meta-tag. Scanning summaries really saves you time if your search returns more than a few items.

Get More Hits By Understanding

Search Engines

Knowing just the little bit above can give you ideas of how to give your page more exposure.

Hustle for Links

Most software agents find your site by links from other pages. Even if you have sent in your URL, your site can be indexed longer and ranked higher in search results if many links lead to your site. One of my sites that couldn't show up in the most casual search got most of its hits from links on other sites. Links can be crucial in achieving good exposure.

Use Titles Early In the Alphabet

All engines that I used displayed results with equal scores in alphabetical order.

Submit Your URL to Multi-Database Pages

It is best to use a multiple-database submission service such as SubmitIt! to save you the time of contacting each search service separately. Remember, it takes 6-8 weeks to become indexed.

Control Your Page's Summary

You can use the meta tag name="description" to stand out in search results. Appear in search summaries as "Experienced Web service, competitive prices" not "Hello and welcome. This page is about."

Search Reverse Engineering

Simulate your audience's search for your page (have all your friends list all the searches they might try), then see what you need to do to come up first on their search engine's results list.

1. Use the meta-tag name="keywords" to put an invisible keyword list at the beginning of your document that would match keywords your audience would use. Most search engines rate your page higher if keywords appear near the beginning.

2. How many times do the keywords appear in the text? It usually demonstrates *good* writing if you don't repeat the same words over and over. However, search engines penalize you for this, usually rating your page higher for repetitions of keywords, inane or not. Some authors combat this by putting yet more keywords at the bottom of their pages in invisible text. Look at the source code for this article, and you'll see what I mean; the words are just in the same color as the background.

SPAMMERS BEWARE

"Spamming" is net-lingo for spreading a lot of junk everywhere; keyword spamming is putting hidden keywords a huge number of times in your document just so yours will be rated higher by search engines.

1. Search engines typically limit you to 25 keywords or less, and one I know of truncates your list when they see an unreasonable number of repetitions.
2. Invisible text at the end of your pages puts blank space there, which looks bad and slows loading. Services which rate pages will enjoy marking

you down for this.

Responsible Keyword Use: If an important keyword doesn't appear at least four times in your document, I hereby give you the right to add invisible text until it appears a maximum of five times.

Comments are welcome

Revised: May 20, 1998

URL: http://webreference.com/content/search/how.html

Search Engines

III. Getting the Most Out of Your Search Engine

Search Engine Features

Web location services typically specialize in one of the following: their search you specify a search and how the results are presented), the size of their datab catalog service. Most engines deliver too many matches in a casual search, so factor in their usefulness is the quality of their search tools. Every search engi nice GUI interface that allowed one to type words into their form, such as "(bu cheeseburger) or (pizza AND pepperoni)." They also allowed one to form Boo (except Hotbot as of 7/1/96, which promises to install this feature later), i. e. th the user to specify combinations of words. In Alta Vista and Lycos, one does t a "+" or a "-" sign before each word, or in Alta Vista you can choose to use the syntax Boolean "advanced search." This advanced search was by far the harde also the one most completely in the user's control (except for OpenText). In m engines, you just use the words AND, NOT, and OR to get Boolean logic.

By far the best service for carefully specifying a search was Open Text. This f menus, making a complex Boolean search fast and easy. Best of all, this servic to specify that you want to search only titles or URLs. But then there's Alta Vi known "keyword" search syntax, now as powerful as OpenText, but not as eas can constrain a search to phrases in anchors, pages from a specific host, image text, document titles, or URLs using this feature with the syntax keyword:sear There is an additional set of keywords just for searching Usenet. (To my know Vista's keywords were undocumented before 7/19/96, so tell your friends you first!)

Which Search Page Should I Use When, and How?		
Use . . .	If You . . .	Using the Feat
Lycos	have no good ideas for specific	best test results for b

	search strategies	terms
" "	want to find someone's e-mail	People Finder.
Magellan	have more than one broad search word, or can't pick a site from Lycos' summaries.	best available results
" "	want interactive news/ want details on today's headlines.	news with links to re
OpenText	want to search only document title or perform complex searches	title search specifica advanced search inte
Alta Vista	are hunting for an image	image:search_word :
" "	want to find all the links to your page	+link:your_site -url:y syntax.
Yahoo!	want the best national and international news	Reuters world headli
" "	want a dictionary or other reference source	Dictionaries or Refe Libraries.

What could really make engines with large data bases shine, however, would b improvement in the way they rank and present results. All engines I tested had schemes that were not well documented, based on how many times your searc mentioned, whether or not they appeared early in the document, whether or no appeared close together, and how many search terms were matched. I did not f ranking schemes very useful, as relevant and irrelevant pages frequently had th scores.

Catalogs

I have only been disappointed by catalog services. In practice, they seem to aim for the lowest common denominator, and reflect very little thought to how and when they might be useful instead of search engines. All the ones I tested were directed toward novices and favored popular commercial sites. I would have thought they would be very good for finding software at least, but this was not the case. See the example below

Useful Non-Search G

E-mail address books:
Most engines allow you for someone's name if y "John Q. Webhead", bu to be careful about exac use of initials, etc.

News Services:

trying to find Web server related software.

Yahoo! has the best nev
humble opinion, as they
Reuters international ne
headlines. Most other n
ultra-brief summaries w
like "MacPaper."

Advanced or Boolean Queries

Making queries very carefully in Boolean terms to narrow a search rarely produces useful results for me (but see below). In practice, other ways of specifying a search besides detailed logic are much more useful. Specification of exact vs. approxi specification that search terms must appear as section headings or URLs, using keywords, and just specifying the language of the document would have been in all of my search examples.

Example: Eliminating Unwanted Matches

The exception to this is the AND NOT operator - it is essential to exclude unw close matches when they outnumber the desired matches. An example of wher operator is given by the problem of finding information on growing apples, be be deluged by information on Apple computers. With enough work, you can s apples with stems, not cords, but it isn't easy. Using Alta Vista, "+apple -mac* soft* -hard* -vendor" got me information on the Payson-Santaquin apple farm a federal apple agriculture database on the first page of results.

Useful Search Features

- **Find Images to Steal (Alta Vista)**

I bet you will all use this at one time or another, so I insist you credit this art webreference.com for this goodie: With Alta Vista, you can limit your searc titles by using the format:

image:title_string

This was the only way I could find a useful picture of a nose for a physician' had searched through jillions of clip art pages, and even contacted graphic a they couldn't come up with anything as good as I found for free! USE THIS.

Try it now (replace ansel with your choice of image search string):

Alta Vista Search:
image:ansel
Search the Web

- **Search for Strings in Titles (Alta Vista, OpenText)** for faster results.

If applicable, this kind of search eliminates chaff by sticking to the pages tha your subject, not ones that just mention a lexically related word. Use the syn

title:search_string

in Alta Vista, or just use the simple pull-down menus in OpenText's "advanc mode."

- **Find the Links to Your Own Site (Alta Vista)**

Alta Vista claims that you can get all the links to your own site by searching keyword construction: +link:http://mysite.com/ -host:mysite in the Simple q ...I found that the most important link to one of my sites was missing from tl so I was not impressed; however, my editor swears by this. Try it now (repla webreference below with your site name):

Alta Vista Search:
+link:http://www.webreference.com -host:webreference
Search the Web

- **Find the Number of Links to Your Own Site (Alta Vista)**

For a more accurate estimate of the actual number of links to your site (or ba use Alta Vista's advanced search, and display the results as a "count only." T method will give you links, but approximates their number, this method mor estimates the number of backlinks. Try it now (replace webreference below site name) ABK-12-29-96:

Search the Web **and Display the Results** as a Count only
Selection Criteria: Please use Advanced Syntax (**AND, OR, NOT, NEAR**).
(link:http://www.webreference.com OR link:http://webreference
OR link:http://www.webreference.com/ OR
link:http://webreference.com/ AND NOT host:webreference
Submit Advanced Query

Which is the Best Search Engine?

(It's not just how big your data base is, it's how you use it.)

To decide which search engine I would choose as the best, I decided that nothi results would count. Previous articles have emphasized quantified measures fc database sizes, but I found these had little relevance for the best performance i searches. By now, all engines have great hardware and fast net links, and none significant delay time to work on your search or return the results. Instead, I ju with a few topics that represented, I felt, tough but typical problems encounter who work on the net: First, I tried a search with "background noise", a topic w closely related but unwanted information exists. Next, I tried a search for som obscure. Finally, I tried a search for keywords which overlapped with a very, v search keyword. I defined a search as successful only if the desired or relevant returned on the first page of results.

Example - Search Terms Which Yield Too Many Matches

For the first type of search, I wanted to find a copy of Wusage to download, fr

that lets you keep track of how often your server or a specific page is accessed tool for HTML developers. This site is hard to find because output files are pr program on every machine running it that have the string "wusage" in their titl When I simply typed "wusage" into search page forms, Infoseek and Lycos w engines to find the *free version* of the software I wanted. (Note I gave no credi the version for sale. A careful search of the sale version's page, did *not* produc the free version's download site.) Infoseek's summaries were very poor, howe matches had to be checked.

Always Search As Specifically As Possible

Most engines failed to find their quarry because the search was too broad. Aft the engine supposed to know I want the free version? After spending a long tir the *exact* name of what I wanted, "wusage 3.2", Infoseek, Excite, Magellan, ar found the site I was interested in. Alta Vista, Hotbot, and OpenText yielded nc interest on their first page. Magellan came out the clear winner on this search, summary was by far the best. (Asking Alta Vista to display a detailed version didn't change things at all!) Infoseek and Excite performed well, but Lycos list older version of wusage (2.4) first.

Think About Search Terms

It eventually occurred to me to search for "wusage AND free" to find the free wusage. In some sense, Lycos was the winner this time because the free versic match listed; however, its summary was not very useful. While it did a better j Infoseek, it didn't tell me whether each site was relevant or not. Magellan's res very good, as it included a link leading to the software on the first page of mat with an excellent summary. Yahoo and Alta Vista also found it, but all these e the fee version higher than the free version. OpenText did very well here, but advanced search mode where it was possible to specify that wusage must be in "free" could be anywhere in the text. Wusage3.2 was listed as the second of or - no digging here! Excite failed to find the site at all, and HotBot found only 1 statistics of a server in Omaha.

Curiously, a search for "download wusage" did not improve the results over th searches for any of the search engines! (It may be time for rudimentary *standa* categories to be used on the Web: e.g. this is a download archive, this is an inf site, this is an authoritative site, etc.) The lesson here may just be "if at first yo succeed..."

Catalogs

Catalogs were not helpful. Yahoo!, under computers/software had nothing wh for wusage: no http, no HTML, no wusage, not even servers. In Excite!, under computing/www/web ware, three more clicks got me to wusage, but -surprise! get to the free version. See why you don't want anyone *else* filtering your info

The lessons from this search, which I have found repeated in other searches, a "Examples: Summary . . ." box below.

Example - Finding The Really Obscure

For this example, let's try to find out how to care for a "tegu", a South American lizard that is only moderately popular even among lizard enthusiasts. (If that's not an adequate example of obscure information, I don't know what is.) I know that a page exists called "TEGU INTRO" at http://www.concentric.net/~tegu/tegu.html, but we will simulate a blind search here. This search was full of surprises.

First I began by just searching for the string "tegu." Infoseek's first match was a tegu page I did NOT know about! Still, the one I wanted was not listed on the first page. Excite yielded nothing about tegus, only information on a vaguely related reptile, the "dwarf tegu." A search on the string "tegu care" yielded nothing relevant. (A search on their handy Usenet database did find the old tegu article I was looking for, three weeks old, which was no longer on my local news server. Other engines found this as well.) Lycos came up with the URL Infoseek found, plus two more, however, the additional listings were only pictures, not information. Searching for the string "tegu care" got nothing. Alta Vista found nothing useful either way, just ads for lizard food. OpenText found nothing, even when I searched for "tegu lizard." Hotbot found a picture of a tegu with "tegu care," but it did not return any relevant information with any search.

None of the searches I tried came up the URL I knew about. The lesson here is that you can really find new things on the Web with search engines, but if you need to find a specific page, it will always be a crap shoot. Advanced searches yielded nothing more with any engine ("tegu in title AND (care or lizard)", etc.) Some way to require that the searches were only among English language documents would have been much more helpful. Some northern-European sounding language apparently has the word tegu in it, not referring to a lizard, and many foreign language pages fouled my results on some engines. Another feature that would really

Examples Summary: I
Improve Your Sea

The most valuable search to
specific information
on a search. (In the sea
wusage, I had no proble
knew that version 3.2 w
needed.)

Think about your search te
next most important search
Obviously, since I want
version of wusage in th
should have searched fo
AND wusage"; I got no
just "wusage" with mos

Good site summaries save y
saving you surfing
Use Magellan or OpenT
possible. To research th
above, I had to pour thr
dozens of pages. Only M
summaries really gave
confidence that I did no
check every site.

Specify a "title only" search
applicable
Title only searches are
only with OpenText an
Vista. In the examples,
more practical results th
up with lots of search w
help pages suggest) or t
forming logically comp
queries (as one might th
Adding more search wo
the results above worse
A Boolean search also
better, e.g.. "wusage AN
download)" yielded not
Alta Vista.

Searches Can Yield New Inf
but they are never complete
None of my searches ev
the good page on tegu

have made a difference would be a filter for sales pages -- most of the mentions of tegu on the net are ads for "Monitor and Tegu Food", containing no care information. As expected, Yahoo! and Excite! Catalogs we as well.

know exists.

Example - Selectivity: Apple Trees NOT Apple Computers

There are gobs of stuff on the net about Apple Computers, but what about gro trees? Surprisingly, this search was very easy! apple* alone always yielded lot about the computers, and one often had to add as many as five excluded terms vendor* -hard* -soft* -comp* -mac*) before receiving any matches for apples Surprisingly, however, just apple* tree* usually yielded detailed information c apple trees on the first page of results. The poorer results required one to incre command to apple* tree* grow*.

And The Winner Is. . .

I don't really want to pick a winner. . . All right, if you insist: The "Search Tes table, below, lists the engines in order of their ranking. **Lycos** is therefore the weight search engine champion of the universe, based on the tests above. How this is missing the point. As shown in the table, "Which Search Page . . . ?", ab should choose different engines for different tasks. None of the engines tested limit their searches to images except for Alta Vista. This engine must therefor best *one* for graphics designers if they are allowed to use only one, but for mo purposes, the user will have to wade through the mountains of chaff and drek t they want. It is more beneficial to use different engines for different tasks; at n few are required.

Search Engine Test Results

Engine	"One Item Among Many Related Pages" Test	"Obscure Item" Test	"Selectivty: Apple Trees Not Computers" Test	Comn
Lycos	Found item with broad search word and exact name. Found item first on results list with two search terms.	Found unknown item, but not known item.	Just apple$ tree$ yielded good results.	Retur releva in the requir time t match Magel
	Found item with broad			

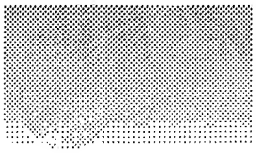
Infoseek	search word and exact name. Found item with two search terms.	Found unknown item, but not known item.	Just apple$ tree$ yielded good results.	Poor S
OpenText	Found wusage in title search	Found Nothing.	Good results with 2 or 3 terms, most useful with 3 terms due to superior summaries.	Abilit title s useful friend Summ good.
Alta Vista	Failed with approximate and exact words. Found item low on first page with two search terms.	Found nothing	Good results with apple* tree* grow*.	Keyw search image are ve other
Magellan	Found with exact name. Found item low on first page with two search terms.	Found nothing	Required three search terms: apple* tree* grow*	Super summ alway surf ti
Excite	Found with exact name, failed with two word search.	Found nothing.	Required third search term: apple* tree* grow*, even then irrelevant results were first.	. . .
HotBot	Failed all searches	Failed all searches	Found only images, and did worse when grow* was added!!!	Poore (exclu catalo
Excite!				

Catalog (not engine)	Failed all searches	Failed all searches	Failed all searches	Catalc useful
Yahoo! Catalog (not engine)	Failed all searches	Failed all searches	Failed all searches	Catalc useful

Comments are welcome

Revised: Dec. 29, 1996

URL: http://webreference.com/content/search/features.html



Search Engines

I. Background

The Web evolved beyond FTP archives not just by becoming a graphically rich multi-media world, but by evolving tools which made it possible to find and access this richness. Oldsters like this author remember that before browsers there was WAIS (released 1991), and the XWAIS version provided a user-friendly GUI way to find information. However, this system required servers to organize information according to a specific format. GOPHER, another information serving system with some user-friendliness, was released the same year. One of the earliest search engines like those today, Lycos, began in the spring of 1994 when John Leavitt's spider (see below) was linked to an indexing program by Michael Mauldin. Yahoo!, a catalog, became available the same year. Compare this to the appearance of NCSA Mosaic in 1993 and Netscape in 1994.

Today there are a score or more of "Web location services." A search engine proper is a database and the tools to generate that database and search it; a catalog is an organizational method and related database plus the tools for generating it. There are sites out there, however, that try to be a complete front end for the Internet. They provide news, libraries, dictionaries, and other resources that are not just a search engine or a catalog, and some of these can be really useful. Yahoo!, for example, emphasizes cataloging, while others such as Alta Vista or Excite emphasize providing the largest search database. Some Web location services do not own any of their search engine technology - other services are their main thrust. Companies such as Inktomi (after a native American word for spider) provide the search technology. These Web location services have put amazing power into every user's hands, making life much better for all of us. . . . and it's all free, right?

. . . Maybe not. It is rumored that these information companies might increase their revenues by selling information - information

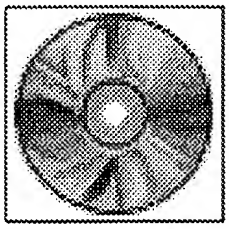
about *you*. After you use a search engine and find a page with mutual fund quotes, you might find yourself suddenly receiving e-mail advertising investments. Think this is a coincidence? Think again. The investment company could have paid a search engine for your e-mail address. The sale of such information is not advertised at this time, however, there is an existing protocol for servers to ask a user's browser for such information, routinely entered during set-up. Get scared about your privacy by checking out the anonymizer snoop page. For best results, search for the anonymizer snoop page, "I can see you", then go to it from your search engine (you'll see what I mean). For now, let's stick to the practical aspects of search engines, catalogs, and Web location services.

Comments are welcome

Revised: August 28, 1996

URL: http://webreference.com/content/search/bkground.html

service bureaus is presented.">



Indexing Digital Documents-- It's NOT an Option

Pay Now or Pay (More) Later

Taking Stock | **Model** | **Glossary** | **Bibliography** | **Credits**

Abstract

Conversion from paper-based filing to an electronic document management system (EDMS) requires significant planning. Indexing digital documents is not optional. This paper distinguishes between field-based and full-text indexing and recommends a combination of the two. Tangible and intangible organizational benefits of indexing digital documents are outlined. The various costs associated with indexing are detailed, and specific price information from service bureaus is presented. Recommendations for choosing an EDMS are included, as well as a model for assessing the organization's indexing needs.

Introduction

Organizations have traditionally relied on paper filing systems for document storage and retrieval. However, paper records are extremely difficult to manage because they have to be stored in and retrieved from only one place. Electronic document management systems (EDMSs) solve many of the storage and retrieval problems inherent in paper filing systems while simultaneously reducing business costs. EDMSs manage storage and retrieval of many different types of digital documents, including word processing files, spreadsheets, database files, e-mail, voice mail, scanned images, and Internet/intranet HTML documents.

While EDMSs provide much faster access to and retrieval of documents (which is a financial benefit in itself), the mere availability of a new technology does not justify its acquisition. The real measure of value "should not be how much faster you are able to respond to a situation with new technology, but rather what value is added to the business process through faster response" (Koulopoulos, 1995). Effective indexing can add value to the organization far beyond mere speed of retrieval by enabling users to retrieve documents in many different ways. Think of business records as part of a hierarchy of "containers" which include Folder, Section, Document, and Page. A folder can have many sections, and sections can contain many documents, and documents can consist of many pages. Yet traditional paper-based filing systems require users to retrieve all information at the "Folder" level of the hierarchy. By contrast, EDMSs allow information to be retrieved at many levels. This retrieval is built on indexing, the bedrock of EDMSs. The accurate and consistent indexing of digital records is absolutely critical to the success of the organization.

So what do you need to know about indexing to increase your document retrieval efficiency and save money? There are many factors which affect indexing needs. First, an understanding of the two basic types of indexing is needed.

Types of Indexing

Indexing can be field-based, full-text, or a combination of the two. Index field data make unique identification of documents possible. For example, the United States Department of Defense is considering the user of a pair of index fields as unique identifier: creation date/time and creator ID. Adding other indexing fields provides additional, controlled ways to access individual records or groups of similar records. Retrieval from index fields is consistent and accurate because it is based on a controlled search vocabulary. Ideally, field indexing is performed at the point when business documents are created. Some field indexing can be done automatically (more on this in the costs analysis), but human indexers are also required.

Full-text indexes are created automatically. Computer software reads every word of every document in a database and creates an inve rted index of words and their locations in the database. End-users can search the database using any words they want to (this is called "natural language"); the computer will find every match between the search term(s) and the text of the documents. Full-text searching makes it easy to locate documents when users are not exactly sure what they need, but it also finds a high number of irrelevant items (for example, Internet search engines are based on full-text indexes). The organization pays for time employees spend browsing through irrelevant documents (or "misses") to find the relevant ones ("hits"). In the interest of quick and accurate retrieval, some field-based indexing is recommended. Indexing digital documents exclusively with full-text indexes is *not recommended.*

All organizations benefit from some combination of field-based and full-text indexing, but determining what particular combination is most beneficial to a given organization is a very complicated process. Before you choose an EDMS to manage your digital documents, your indexing needs should be weighed against the benefits and costs of indexing. Indexing is *not an option* with EDMSs--the documents have to be indexed in some way. Different EDMSs offer different types of indexing, and the organization should be aware of their capabilities. Organizations have different indexing needs because their documents and their users vary. This article details the benefits and costs of indexing digital documents and includes a model for assessing the indexing needs of the organization.

Organizational Benefits of Indexing

Indexing digital documents produces both tangible and intangible benefits to the organization. Tangible benefits include financial, legal, employee, and value-added benefits. Intangible benefits include less concrete measures of success, such as improved perception of the organization by both employees and customers. Combined tangible and intangible benefits result in financial gain for the organization through increased employee productivity, customer service, and competitive advantage in the marketplace.

Financial Benefits:

- Increased production. The speed of many routine office procedures (such as production of statistical reports, records management tasks, access to and retrieval of digital documents, etc.) is increased.

- Decreased future staff requirements. Increases in production can be handled by current staff.

- Increased access to current information. Quick and accurate updates of indexes throughout the organization decreases information retrieval time and increases accuracy of information.

- Improved customer service. Prompt, accurate information retrieval increases repeat revenue for the organization.

- Decreases in human filing mistakes. Large legal practices often spend 8 or more hours to locate misfiled documents (Socha , 1996).

CASE STUDY

A legal firm using an image management system found that their cases could be handled by 2.5 fewer temporary full-time clerks than before they implemented the system. With the previous paper-based system, clerks spent large amounts of time retrieving documents identified in database searches, photocopying the documents, delivering the copies to attorneys and legal assistants, and refiling the originals. The clerks also spent considerable time searching for misfiled originals. 2.5 clerks earning $14/hour for 160 hours/month over 14 months would have cost the firm $78,400 (Socha, 1996).

Legal Benefits:

- Litigation protection. In a lawsuit, records need to be produced very quickly. An indexing system that can identify and retrieve documents needed for litigation can pay for itself if a single multi-million dollar lawsuit is avoided.

- Response to Rule 26. A new law requires parties involved in a federal lawsuit to identify and produce relevant records within 85 days of the beginning of the litigation (Skupsky, 1995). Quick and accurate retrieval of records is required.

- Records retention compliance. Federal, state and local governments regulate record retention periods for organizations. There are over 10,000 federal recordkeeping laws alone (Skupsky, 1989). Good indexing systems include indexing fields related to retention (such as creation date, retention period, and disposition date).

CASE STUDY

As a result of Rule 26, courts will probably require each party involved in a lawsuit to make a full disclosure of their records in the early stages of the case. Sanctions will follow for parties which fail to produce relevant information. Disorganization of records will not excuse parties from compliance. For instance, in United States v. ABC Sales & Service, the court concluded that "'a business that generates millions of files cannot frustrate discovery by creating an inadequate filing system so that individual files cannot readily be located'" (Skupsky, 1995).

Employee Benefits:

- Currency of business information. New documents can be added to the indexing system quickly, and if documents are indexed when they are created, all users can access them immediately. Employees can do their jobs better.

- Document version control. Indexing digital documents makes it possible to control which version of a document users can access. Employees don't waste time working on outdated documents, or updating a version that's already been revised.

- Remote access. An organization-wide standard indexing language allows authorized users to retrieve documents from anywhere in the world. Employees don't have to take their whole office with them when they travel.

- Simultaneous access. Employees can share a document if it is indexed properly and retrieved from a computer network. The "file folder" is never missing from the file cabinet. Hard copy production and distribution are also eliminated.

- Decreased training time. New employees become quickly and fully productive in the organization.

CASE STUDY

When the U.S. Patent and Trademark Office (PTO) implemented a new imaging system, its most noticeable benefits involved customer service and employee training. The PTO Commissioner said that new patent examiners learned the business much faster because of the indexing system. The old manual indexing system required about 12 years to master; new examiners trained on the imaging system were up to speed in just a few months (Koulopoulos, 1995).

Value-Added Benefits:

- Customer service improvements. Organizations that provide high levels of service will gain customer loyalty and increase business.

- Competitive advantage. Organizations that can retrieve information quickly and accurately will be able to accomplish more during the work week. Time is money, and indexing saves time.

- Perceived excellence. Companies that project an image of excellence will attract more clients and better employees.

CASE STUDY

Pharmaceutical giant Glaxo implemented an EDMS and saved over $1 million per year associated with search and retrieval time. However, financial benefits were not the most valuable benefits realized. Each New Drug Application process requires about 50,000 pages of data preparation and documentation; the EDMS and its indexing system allowed Glaxo to prepare this documentation and receive clearance from the Food and Drug Administration much more quickly than before. Thus, EDMS implementation enabled Glaxo to collapse their business cycle and get their product to market sooner than their competitors (Perkins, 19??).

Costs of Indexing digital documents

How much will it cost to index your digital documents? One vendor quickly replied, "How much do you have?" But that answer is neither realistic nor helpful. Companies contemplating development of an indexing system for digital documents want to spend as little as possible to obtain a retrieval system that is needed to conduct business. More specifically, they want a system that provides quick and accurate access to frequently-retrieved information and reliable (but not necessarily fast) access to infrequently-retrieved information.

Because the types of business documents which meet these criteria in different organizations vary so widely, it is obvious that there is no one "best" indexing scheme. One size will never fit all. Therefore, indexing costs will be detailed in two ways: 1) factors that affect the cost of indexing, and 2) cost information reported in published studies (see Table 1).

Factors That Affect the Cost of Indexing:

One of the first decisions which must be made is whether documents not currently in digital form will be converted. A paper or mICRofilm document is converted to digital format by scanning it into a computer; OCR/ICR (optical character recognition/intelligent character recognition) software may then be used to convert the document to ASCII text (Thiel, 1992). Documents can be indexed before or after they are scanned. Spencer (1996) estimates that the true cost of batch scanning 10,000 documents is about $.09/page before indexing costs are included. Thus, undertaking a large document conversion project can be costly. DocuCon, a full-service document conversion firm, comments that at least 20% of the documents to be scanned will require special handling (because of size or condition) and that rated equipment speeds are not reliable guides to how long jobs will actually take; special conditions like these further increase the cost of document conversion (Cullen, 1991) . Other factors which affect the costs of indexing include the cost of keying index field data, technological costs, retrieval costs, and costs of updating.

Manual field indexing of digital documents can be performed when the documents are created or when they are stored. For example, electronic document processing systems often require that employees who produce letters and reports using word processing/spreadsheet software fill some index fields when the document is saved. Although the time required to index a single word-processed document is small, the individuals who do this indexing may be highly paid, which increases the overall cost of indexing digital documents. The most variable (and often the highest) cost associated with indexing is labor.

Indexing cost can be minimized by searching for ways to fill index fields from information already contained in existing corporate databases. If manual entry of a customer number allows the system to automatically access name, address, or zipcode, a great deal of manual keying time may be eliminated (Devlin, 1996). Barcoding is a new and cost-effective way to quickly and accurately identify batches of document types or individual documents (Spencer, 1994). For example, if a type of business form is preprinted with a bar code that identifies what type of document it is, the EDMS can automatically populate the "document type" indexing field when the document is scanned and OCRed. No one has to key the document type, which decreases cost.

The number of index fields used to identify a particular document is a significant cost factor, especially when indexing is performed manually. A study of indexing projects showed that the average number of index fields is 8-12 (Cisco, 1993). However, an ANSI Technical Report prepared by the Association for Information and Image Management International suggests 50 possible index fields which might be used with electronic image management systems (AIIM, 1995). If the average field contains 12-20 characters, the cost difference between manually keying each additional field must be considered.

Sometimes the cost of indexing documents can be reduced or eliminated by using full text retrieval systems which create an additional file (usually called an inverted file) in which each non-trivial word is listed with a locator key (Thiel, 1992). full text retrieval systems also allow users to construct search queries in their own words, rather than having to conform to the restraints of pre-selected terms (Fidel, 1994). However, full-text systems often return an unacceptably low number of relevant documents, fewer than 20% in one study (Blair & Maron, 1985). Some organizations will be unable to afford the cost of not finding relevant documents every time they look for them.

Technological Costs

Although most organizations are already computerized and the cost of adding computer capability and memory storage is becoming increasingly economical, there still remain technological cost implications in choosing indexing systems. The size of the index itself must be considered. Inverted files (used by full text retrieval systems) are often very large, sometimes requiring more storage space than the documents which they index (Thiel, 1992). Timely document retrieval may require faster processing

speeds than the organization presently supports. And if documents are being shared by many users, local area networks may have to be installed.

The cost of data migration (which includes index migration) must also be considered. Organizations should appoint an information management professional to administer data migration and indexing so that documents remain accessible as technological change occurs. Many organizations already own systems that contain non-standard or proprietary software which makes integration and migration difficult. Planning for future technological change now will save costs later.

Retrieval Costs

If minimizing the costs of indexing documents ultimately increases the cost of retrieval, it may be false economy. Kind and Eppendahl (1992) suggest a number of questions which must be asked about document retrieval, including who performs searches, how frequently items are needed, how long each search takes, how quickly the information must be made available, and how often a needed document cannot be found. Answers to such questions have cost implications which must be considered when designing an indexing system. For example, an inexpensive indexing system will require more search and retrieval time than a more expensive one. Can you afford to have your highly-paid employees spend time searching for and retriving documents? If you don't invest in the indexing system, you will pay for it (and pay more for it) in retrieval.

Another retrieval cost involves training employees to use the system. The more complicated the indexing scheme, the more time and training will be required before users feel comfortable and confident about their ability to access the information they need.

Cost of Updating

Two different Kinds of updating costs must be considered. First is updating the documents in the system. If most documents exist in only one version, it may be economically feasible to simply start indexing over each time a document is revised, essentially giving it a new identity. However, if documents are frequently revised or modified, the organization may need to identify the most recent or official version of a document. Additional indexing fields may be needed to ensure that multiple users all have access to the latest version.

The index itself must be kept current and updated. Griffiths and King (1993) survey 16 organizations and suggest that direct costs of an "index maintenance" project average $.29 per document (the project included creation and addition of new terms, removal of obsolete terms, and authority and location control work). Index maintenance may cost more that the original cost of indexing documents. Time and effort spent on initial index design may eliminate costly projects to correct or update after the system is in place.

Cost of Indexing

Table 1 shows examples of costs and ranges found in published studies of indexing projects. Koulopoulos(1995) reports that the time spent designing a typical system is divided among field identification and data standardization (20%), data entry (20%), and system correction and fine-tuning (60%). Initial purchase of digital imaging systems with capacity to process and store 300,000 to 3 million pages per year costs $.15 to $.25 per page, depending on use.

Costs reported by companies indexing their documents in-house range from $.12 to $.20 per page

(Cisco, 1993). Typical service bureau charges currently range from $.15 to $.30 per page for scanning and indexing (it is not clear how many index fields would be included).

Conclusion

So how few index fields can your organization get by on? You need at least two fields to ensure data retrieval--one uniquely identifies each document, and another provides an alternate pathway in case the first one fails. W. Wiggins of DocuCon recommends indexing a unique identifier and the document type for each document (personal communication, August 3, 1996) . You need additional fields to manage records retention and disposal. You also need to index processing information about the software and hardware used to create each document so that data can be properly migrated when necessary. The Association For Information and Image Management (AIIM) identifies 30 possible processing information fields and 20 possible retrieval information fields (1995). The United States Department of Defense uses 22 records management fields to index their documents (Prescott, Underwood, & Kindl, 1995).

Answering the questions in "Taking Stock of Your Company's Indexing Needs: Full-Text, Field or a Combination?" will help you identify what sort of data needs to be stored in index fields. Obviously, we cannot recommend a minimum number of indexing fields needed to effectively retrieve business documents. Each organization has unique requirements that should be thoroughly studied before implementing an indexing system.

Document Demographics

What requirements do your documents fulfill?

- Business purposes (to make payroll, pay bills, write reports, serve customers)
- Legal purposes (to prepare for litigation, audits, regulatory reporting)
- Records management purposes (to manage retention, disposition, vital records protection)
- Archival purposes (to conduct longitudinal studies, genealogical research)
- What is the condition of the documents and the information contained on the documents?
- (Are the documents legible enough for more than 90% to be OCRed? Are the documents brittle, torn, stained, or skewed?)
- Do you have documents that are created electronically? (Example: word-processed documents, IRS income tax returns submitted electronically)
- What indexing data are already available in existing corporate databases? How accurate, complete, and consistent are the available data?

User Demographics

- Who are the primary users? How do they ask for the documents?

- Who are the secondary users (such as outside auditors, strategic business partners, customers)? How do they ask for the documents?

Benchmarking

- How do other organizations in your industry index documents?

- How do other organizations outside your industry index documents?

- Are there industry standards for indexing?

RECOMMENDATIONS

- The purpose of indexing is retrieval. If documents cannot be found, they may as well not exist.

- Think about ways people in different parts of the organization might want to retrieve the same information. You may have secondary users with different indexing requirements than the primary users. It is better to purchase and implement an indexing system that fits your users than to try to change your users to fit the system.

- Make sure that your users are trained to use the indexing system that you have implemented. Ask your vendor if their system includes the costs of training your employees.

- More is not necessarily better. Six to ten index fields may be as useful to your users as fifty, at much lower cost. However, one or two extra fields may be worth the investment if they help avoid litigation or simplify compliance with regulations. Every organization has different and unique indexing needs.

- Think about how to use information you already have in digital format to avoid additional indexing and re-keying. You may be able to automatically populate some index fields.

- Use a full-text index in addition to indexing fields to index documents that are created electronically (such as word-processed documents, online application forms, etc.). To further reduce costs, consider introducing more digital document creation in the future.

- Include controlled vocabulary indexing fields to standardize indexing terminology. You don't want different departments calling the same kind of document different things (such as "bill," "voucher," and "invoice").

- Work with a reliable vendor who uses non-proprietary programming language.

COMPLEXITY OF INDEXING NEEDS

How many indexing fields do you need?

Document Demographics

	... each document fulfills few organizational requirements, then:	... each document fulfills many different organizational requirements, then:
If users ask for documents in similar ways, and ...	Few indexing fields are needed	Medium number of indexing fields is needed
If users ask for documents in different ways, and ...	Medium number of indexing fields is needed	Many indexing fields are needed

Glossary

ANSI (American National Standards Institute): an institution which develops and publishes standards for use within the United States.

Automated Indexing: computerized indexing which doesn't require human decision-making or data entry. Automatic indexing software populates index fields by reading information from bar codes or scanning digital documents which have undergone OCR conversion.

Barcode: a sequence of machine-readable lines of varying widths which contain data. Barcodes can be used to facilitate automatic indexing. For example, if standard business forms (such as invoices) are preprinted with a barcode which indicates that the form is an invoice, an indexing system can automatically populate the "document type" field after the paper form is scanned and OCRed. Barcodes also survive fax transmission intact.

Batch Processing: a technique by which items to be processed are collected into groups prior to processing.

Coding: See Indexing

Controlled Vocabulary: set of rules for choosing words and phrases to be used in an indexing system, along with the list of approved or allowed words to be used in the system.

Data dictionary : organized collection of information about data. The data dictionary compiles data about data, or metadata. A data dictionary is an automatic component of most database management systems.

Data Element : a unit of data that is considered to be indivisible. Data elements are the building blocks

for all data processing systems. Examples: document type, creation date, disposition date, Social Security Number, etc.

Descriptor : See Field Data

Digital Document : document which exists in electronic form inside a computer system.

Distillation : the process of elminating, summarizing, or in some other way reducing a body of information to its essential components.

Document : 1. any format which contains information. Documents may be word-processing files, e-mail messages, spreadsheets, database tables, voice mail or other audio recordings, faxes, business forms, images, information captured from the Internet, and so forth. Documents are sometimes called "records." 2. According to ANSI/AIIM TR40-1995, a collection of zero or more pages that are related, linked, or bound to each other in some way appropriate to the application. In an electronic image management system, the provision of a zero-page document allows the creation of a document entity prior to capturing and linking its page(s)..

Document Classes : types of documents which require similar indexing fields. Examples of document classes: invoices, contracts, timesheets, e-mail messages, and so forth. Often called "document types."

Document Life Cycle : the period which includes creation, maintenance, use, and ultimate disposition (destruction) of a document. The records manager needs to know the life cycle of every document in the organization.

EDM (Electronic Data Management): application of technology to save paper, speed up communications, and increase the productivity of business processes.

EIM (Electronic Image Management): system which organizes information in all formats for use throughout its life cycle.

Field Data : the retrievable information which follows the field name . Example: for the field name

document type, the field data might be *invoice* or a code which represents *invoice*. The field data concept is associated with many terms, including indexing value, term, and structured or unstructured data.

Field Name : the name of the field where a specific Kind of information is to be entered. Think of "field name" as a prompt for what Kind of information is stored in the field. Field names must be decided on before any documents are indexed. The information stored in the field is called field data . Example: for the field name *document type*, the field data might be *invoice*. The "field name" concept is associated with many terms, including "index key," "key field," "fixed field," and "indexing field."

Fixed Field : See Field Name

Free Text Searching : See Full-Text Retrieval

Full-Text Indexing : indexing method in which the computer creates an alphabetical inverted index consisting of all words (except stop words) in the document along with pointers (locations) to locate the words in the document. Full-Text indexes are inexpensive to create since humans are not needed to define field names or enter indexing values into those fields.

Full-Text Retrieval : a type of retrieval process that uses an inverted index to retrieve every document that contains the word or words in the search parameter. This type of searching requires a powerful search engine and is much slower than retrieval processes based on indexing values. It is also much less accurate because it is not based on standardized search terms. For instance , a search that retrieves all documents containing the word "invoice" will miss those which are designated as "bill" or "voucher." However, full-text retrieval systems initially are cheaper to implement because indexing costs are eliminated. Full-text retrieval is sometimes called "free text searching" or "fuzzy searching." Contrast with keyword retrieval .

Fuzzy Searching : See Full-Text Retrieval

Homonyms : words that are spelled the same but have different meanings. Computers don't recognize homonyms.

ICR (Intelligent Character Recognition): a form of OCR (optical character recognifiton) which uses sophisticated lexical tools. ICR is typically used to convert handwritten material to ASCII text.

Index Key : See Field Name

Indexing : 1. the process of identifying various pieces of information in a document (such as author, document type, creation date, etc.) and then transferring that information into a database for search and retrieval; also called "coding" in the legal profession. 2. the process of analyzing the information content of recorded knowledge and expressing this information content in the language of the indexing system (NFAIS Indexing in Perspective Education Kit) 3. the representation of the results of the analysis of a document by means of a controlled or natural language system

Inverted Index : a computer file in tabular format, in which rows represent documents and columns represent words. Intersections of rows and columns are marked when certain documents contain certain words. At the point of retrieval, the computer scans the entire inverted index for documents which contain the words in the search query.

Islands of Information : corporate information stored in separate and unlinked repositories (such as individual workstations). Storing corporate knowledge in islands of information leads to duplication of effort and difficult (at times even impossible) retrieval.

Key Field : See Field Name

Keyword Retrieval : a type of retrieval process that searches an index with fields to locate documents which contain information related to the search parameter. Contrast with full-text retrieval . Keyword retrieval requires indexing of documents but provides extremely accurate retrieval as long as the indexing is accurate. To guarantee accuracy of indexing, data elements and indexing values should be carefully designed to match the retrieval needs of the document users, and quality control should be part of the indexing process. As one vendor told us, "you get what you pay for in indexing."

Life Cycle : See Document Life Cycle

Mark Sense Code : a method of automatic indexing in which the person responding to a questionnaire or form does so by filling in bubbles or other spaces. A scanner passes over the marks and reads them automatically into the computer, digitizing the responses.

Metadata : data about data. Metadata is information required to document the characteristics of and relationships between information contained within databases (field names, length of field, type of data, etc.). Sometimes called "higher level information" or "processing information."

OCR (Optical Character Recognition): the process of electronically reading digital images (those which have already been scanned) and converting them to text. After OCR conversion, a document is "live," or editable. For instance, users can edit OCRed documents on the computer as if they were word processing documents that they created.

OCR Repair : manual examination and correction of OCR conversion. Some OCR software is capable of flagging documents which it couldn't convert, so that a human is needed to examine and correct only the flagged documents (rather than all of them)..

Ontology : a taxonomy of everything that divides human knowledge (or more commonly, a subset of human knowledge) into a clean set of categories. Example: the Dewey Decimal system.

Page : a page is equivalent to one side of a 2-dimensional sheet of paper, microfilm, transparency, etc. In the case of input media other than paper, a page will be the data in a single image frame (ANSI/AIIM TR40-1995)..

Remote byproduct image capture : the process of reusing scanned images or indexing captured for some other purpose. Typically, digital documents are transmitted to a central collection point and indexing software captures pre-processed information (this information may be housed in a pre-existing database, encoded in bar codes, etc.). "Captured" information doesn't need to be keyed by data entry operators and therefore reduces the cost of indexing. The more byproduct capture a document management system includes, the more cost-efficient it will be (Spencer)..

Retrieval : recovering desired information or data from an organized collection of information.

Retrieval Information : that information necessary for an end-user to retrieve the document after the document has been captured (ANSI/AIIM TR40-1995). Retrieval information may be field names , field data , or a combination.

Single Point of Access : a user-centric information system that provides access to all information through one interface. Information may be housed in databases, word processing files, spreadsheets, e-mail archives, the Internet, voice mail archives, etc. Single point of access is presently a *concept*, not a *reality*.

Spider : a simple computer program that scans the World Wide Web, "crawling" from link to link in search of new sites. The Inktomi internet search engine is a massive spider.

Strategic gain : "an influence which goes beyond meeting immediate operational objectives, and which can positively impact organization structure and/or direction, and therefore performance."

Structured Database Index : a database that has been constructed with fields to receive structured information. Structured information is information about something is known. For example, a field designed to receive a Social Security number must be exactly 9 characters long. A field designed to receive a name should be about 30 characters long to accommodate long names.

Synonyms : words that are spelled differently but have the same meaning. Computers don't recognize

synonyms very well. Example of synonymous terms: invoice, bill, and voucher.

Verification : the process by which data entry is performed twice by one operator, or once by two operators, and the computer verifies that the same data were entered each time. If there is a discrepancy in the data, the computer prompts the operator to enter the data a third time.

Version Control : a method to ensure that the most recent or official copy of a document is the one available for use.

White paper : an authoritative report issues by an organization. Can also refer to an official government report.

Workflow : the amount and flow of work to and from an employee, department, or office. The efficiency of workflow is greatly facilitated by imaging systems which electronically transfer documents from person to person (as opposed to a paper file folder traveling from inbox to inbox). Imaging systems can also transfer electronic documents which are part of the same file to different people and then reassemble the information at a later point.

Bibliography

Acton, Patricia, CRM. July 1986. Indexing is not classifying-and vice versa.Records Management Quarterly: 10-12, 14.

Association for Information and Image Management International. 1995ANSI/AIIM

TR40- 1995 Technical Report for Information and Image Management-Suggested Index fields for Documents in Electronic Image Management (EIM) Environments. Silver Spring: AIIM International.

Bakewell, K.G.B. 1978. Classification and Indexing Practice. London: Clive Bingly.

Blair, David C., and Maron M.E. March 1985. An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System. Communications of the ACM: 289-299.

Borko, Harold, and Bernier, Charles L. 1978. Abstracting Concepts and Methods. New York: Academic Press.

Borko, Harold, and Bernier, Charles L. 1978. Indexing Concepts and Methods. New York: Academic Press.

Chen, Ching-chih. 1984. MICRoUse Directory: Software. Amersfoort, Netherlands: Johan van Halm.

Cisco, Susan. 1993. Indexing Documents for Imaging Systems: A Roadmap to Success. Austin, TX: Marketfinders.

Cleveland, Donald B., and Cleveland, Ana D. 1983. Introduction to Indexing and Abstracting. Littleton, CO: Libraries Unlimited.

Collison, Robert. 1971. Abstracts and Abstracting Services. Santa Barbara,CA: ABC-Clio.

Collison, Robert. 1959. Indexes and Indexing. London: Earnest Bean.

Collison, Robert. 1962. Indexing Books: A Manual of Basic Principles. New York: John deGraff.

Craven, Timothy C. 1986. String Indexing. Orlando, FL: Academic Press.

Cremmins, Edward R. 1982. The Art of Abstracting. Philadelphia, PA: ISI Press.

Cullen, Gerard. 1991. Rules of Thumb for Backfile Conversions on Optical Imaging Systems. Document Image Automation. 11(2):163-165.

De Jaeger, H.K. 1973. Information Storage and Retrieval Without Computer Assistance. ASTRID Series on Information Science, no. 3., Belgium: ASTRID House.

Devlin, Joseph. June 1996. Search and Retrieval the Best Strategies Now;
By Picking the Right Indexing Solutions, Savvy VARs Can Make Big Profits. Imaging Business 2(6): 28, 30, 32, 34.

Fidel, Raya. 1994. User-centric Indexing. Journal of the American Society for Information Science 45 (8): 572-576.

Griffiths, Jose-Marie, and King, Donald W. 1991. Cost Indicators for Selected Records Management Activities
: A Guide to Unit Costing for the Records Manager. Vol. 1. Prairie Village,
KS: Association for Records Managers and Administrators International.

Indexers On Indexing. 1978. Edited by L.M. Harrod. New York: Bowker.

Indexing in Perspective Education Kit. 1979. Edited by Everett H. Brenner. Philadelphia, PA: NFAIS/UNESCO.

Indexing and Abstracting. 1980. Compiled by Hans H. Wellisch. Santa Barbara: ABC Clio.

Kind, Joachim, and Eppendahl, Frank. 1992. The Need for Office Analysis in the Introduction of Electronic Document Management Systems. Document Image Automation. 12(2): 31-35.

Koulopoulos, Thomas M., and Carl Frappaolo. 1995. Electronic Document Management Systems: A Portable Consultant. New York: McGraw-Hill, Inc.: Chapters 1-2 and 5.

Lancaster, F.W. 1991. Indexing and Abstracting in Theory and Practice. London: The Library Association.

Maizell, Robert E., Smith, Julian F. and Singer, T.E.R. 1971. Abstracting Scientific and Technical Literature. New York: Wiley.

Milstead, Jessica L. 1994. Needs For Research in Indexing. Journal of the American Society for Information Science 45(8): 577-582.

Neufeld, Lynne M., Cornog, Martha. 1983. Abstracting & Indexing Career Guide. Philadelphia, PA: National Federation of Abstracting and Information Services.

O'Connor, Brian C. 1996. Explorations in Indexing and Abstracting. Littleton CO: Libraries Unlimited.

Pan, Elizabeth, Dale, Tom, and Beverly, Kevin. 1992. Handbook of Image Storage and Retrieval Systems: Ch. 13.

Penn, Ira. July 1983. Understanding the life cycle Concept of Records Management. ARMA Records Management Quarterly 17: 5-8, 41.

Perkins, S. M. The Business Benefits and Justification: Document Image Processing. Sections 1 and 2.

Phillips, John T. October 1995. Metadata-Information About ELectronic Records. Records Management Quarterly 29(4): 52-55, 73.

Prescott, Capt Daryll R., William Underwood, Ph.D., and LTC Mark Kindl. August 28, 1995. Functional Baseline Requirements and data elements for Records Management Application Software. Atlanta: Army Research Laboratory.

Richmond, P.A. 1981. Introduction to PRECIS for North American Usage. Littleton, CO: Libraries Unlimited.

Rowley, Jennifer E. 1988. Abstracting and Indexing. 2nd ed. London: Clive Bingley.

Savic, Dobrica. October 1995. Automatic Classification of Office Documents: Review of Available Methods and Techniques. Records Management Quarterly 29(4): 3-18.

Skupsky, Donald S., JD, CRM. 1989. Recordkeeping Requirements. Englewood: Information Requirements Clearinghouse: Chapters 1-3.

Skupsky, Donald S., JD, CRM. 1995. Law Records and Information Management: The Court Cases. Englewood: Information Requirements Clearinghouse: Chapter 10.

Socha, Jr., George J. 1996. The Paper Chase: Imaging Can Lead the Pack. Proceedings of the ABA's 10th Annual Techshow. The Convergence of Technology and the Legal Profession: Charting the Next Ten Years: 1-17.

Soergel, Dagobert. 1994. Indexing and Retrieval Performance: The Logical Evidence. Journal of the American Society for Information Science 45 (8): 589-599.

Soergel, D. 1974. Indexing Languages and Thesauri: Construction and Maintenance. Los Angeles: Belville.

Soergel, Dagobert. 1985. Organizing Information: Principles of Data Base and Retrieval Systems. Orlando, FL: Academic Press.

Spencer, Harvey. April 1994. Image Capture for Document Imaging: Where To Do It To Make It Work. Advanced Imaging 9(4): 46-48.

Spencer, Harvey. February 1, 1996. The Recent Evolution Of Data Entry: The Internet,OCR/ICR and Off-Shore Data Entry Can Dramatically Reduce the Cost of Capturing Data From Existing Forms. IW: 37-39.

Steinberg, Steve G. May 1995, Seek and Ye Shall Find (Maybe). Wired: 108-115, 174, 176, 178-180, 182.

Straus, L.J., Shreve, I.M., and Brown, A.L. 1972. Scientific and Technical Libraries, Their Organization and Administration. 2nd ed. New York: Becker & Hayes.

Strong, Karen. 1996. LCRA Uniform Filing Structure (UFS); A White Paper from LCRA Records and Information Management Services (RIMS). Austin: Lower Colorado River Authority.

Thiel, Thomas J. 1992. Automated Indexing of Document Image Management Systems. Document Image Automation. 12(2):43-49.

Townley, Helen M., and Gee, Ralph D. 1980. Thesaurus-Making; Grow Your Own Word-Stock. Boulder, CO: Westview Press.

Wellisch, Hans H. 1991. Indexing from A to Z. Bronx: H.W. Wilson.

Willis, Don. 1994. A Hybrid Systems Approach to Preservation of Printed Materials (Part Two). MICRoform Review 23(1): 18-25.

Worobec. Bruce. October 1994. Enterprise Data One Step at a Time. Database Programming & Design: 64-67.

Cooperating Organizations

Image Scanning of America, 2363 Teller Rd. #127, Newbury Park, CA 91320.
(805)375-0422;fax (805) 375-3062; e-mail: sales@isausa.com
Contact Name: Manuel Bulwa, Director.

searching electronic documents Homepage | Advanced Search

Search using: Ask Jeeves Google

CUSTOM WEB FILTERS HotBot Skins | Prefe

Date: **Before December 15 1999** [Edit this Search]

SPONSORED LINKS (filters not applied)

- Digital **Documents**, LLC ®
  Leading Document Scanning Services Value Pricing & Quality Services!
  www.DigitalDocumentsLLC.com
- Search your **Documents**
  Search engine for **documents**, text, PDFs, emails & more. Free trial!
  www.isys-search.com
- eDrawer $699.95
  SO/HO, SMB and Large Organizations Over 10 000 users and growing
  www.edrawer.com
- Create **Electronic** Forms
  Add content directly into an XML doc or relational DB - Download now
  www.Altova.com/Stylevision
- Easy Document Management
  PaperMaster makes a paperless office a reality. Try it free!
  www.papermaster.net

SI

Free Document Search
Download free 5 000-docur
expiration. For Windows
www.coveo.com

Digital **Documents**
A world class provider for a
document management ne
www.d-docs.com

Documents Searchin
Scan **documents** fast - fin
seconds with full text searc
www.documentlocator.
Place

WEB RESULTS by Ojoo (Showing Results 1 - 10 of 50,900)

1. Government Cd-Roms: A Practical Guide to **Searching Electronic**...
Government Cd-Roms: A Practical Guide to **Searching Electronic Documents** Databases Pop Lists...
queerpopculture.com/entertainment/asinsearch_0887368875/

2. Government Cd-Roms: A Practical Guide to **Searching Electronic**...
Government Cd-Roms: A Practical Guide to **Searching Electronic Documents** Databases sho talkhd.tv...
www.talkhd.tv/amazon/asinsearch_0887368875.html

3. Indexing Digital (**Electronic**) **Documents** -- It's Not an Option; Pay...
Electronic document management systems (EDMSs ... s) and the text of the **documents**. Ful **searching** makes it easy to locate **documents** when...
fiat.gslis.utexas.edu/%7Escisco/inel.html

4. Translating Mathematical Markup into HTML
Translating Mathematical Markup for **Electronic Documents**. Keith Shafer Roger Thompson. A This is done by **searching** for start and end...
www.w3.org/pub/Conferences/WWW4/Papers/177/

5. Creating **Electronic Documents** that Interact with Diagnostic...

Creating **Electronic Documents** that Interact with Diagnostic Software for On-Site Service ...
menu item for **searching** within the...
www.sil.org/sgml/harmison.html

6. Guidelines for Managing **Electronic Documents** in Australian
...time wasted **searching** for **electronic documents** stored without adequate planning for futı
or that have been deleted by mistake due...
www.tomw.net.au/edgcase.html

7. By Subject - Social Sciences - **Electronic** Texts & **Documents**
Off-Campus Access. **Electronic** Texts & **Documents**...
www.lib.washington.edu/subject/socialsci/dr/eltxt.html

8. By Subject - International Studies - **Electronic** Texts & **Documents**
Off-Campus Access. **Electronic** Texts & **Documents** ... 1942-43 New York-Moscow KGB messa
Peregrine Falcon survey in Canada /
www.lib.washington.edu/subject/International/dr/eltxt.html

9. EHRAF via Web
Electronic HRAF (Human Relations Area Files) ... looked for in the text of the **documents** (boc
Rutherford, for assistance in **searching** this...
www.ucalgary.ca/library/subjects/ANTH/ehraf.htm

10. Home: WWW and CGI info
Searching. AltaVista. Lycos. Webcrawler. Yahoo ... SGML and **Electronic Documents**. Online
Library Center. SMGL on the WWW. SGML WWW...
www.vuse.vanderbilt.edu/~drl/home/wwwcgi.html

« Previous | Next »

Search for "**searching electronic documents**" using: Google

search electronic documents Homepage | Advanced Search

Search using: Ask Jeeves Google

CUSTOM WEB FILTERS HotBot Skins | Prefe

Date: **Before December 15 1999** [Edit this Search]

SPONSORED LINKS (filters not applied)

- Digital **Documents**, LLC ®
  Leading Document Scanning Services Value Pricing & Quality Services!
  www.DigitalDocumentsLLC.com
- **Search** your **Documents**
  Search engine for **documents**, text, PDFs, emails & more. Free trial!
  www.isys-search.com
- eDrawer $699.95
  SO/HO, SMB and Large Organizations Over 10 000 users and growing
  www.edrawer.com
- Free Document **Search**
  Download free 5 000-document version. No expiration. For Windows
  www.coveo.com
- Create **Electronic** Forms
  Add content directly into an XML doc or relational DB - Download now
  www.Altova.com/Stylevision

SI

Need Document Storac
Organize your files in Paper of on your desktop.
www.papermaster.net

Digital **Documents**
A world class provider for a document management ne
www.d-docs.com

Documents Search
Scan **documents** fast - fin seconds with full text searc
www.documentlocator.

Place

WEB RESULTS by Ojeeves (Showing Results 1 - 10 of 230,800)

1. ISO 690-2, Bibliographic references to **electronic documents**
Search the ISO Catalogue [from "ISO Online" site] ... 7.9.2. **Electronic documents** spanning one date
www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm

2. ISO 690-2: additional examples of bibliographic references
Additional examples. ISO 690-2 [Bibliographic references - **Electronic documents**] ... **Search** Catalogue [from "ISO Online" site]
www.nlc-bnc.ca/iso/tc46sc9/standard/690-2ex.htm

3. Digital Library - Online Recherche (GBV Online Resources /WebDOC...
Resources Recherche und Download elektronischer Dokumente deutscher WebDOC-Teilnehmer retrivial of **electronic documents** of the...
cosmic.rrz.uni-hamburg.de/docs/webdoc_a.html

4. The **Electronic** Privacy Papers: **Documents** on the Battle for Privacy
Search 80 Bookstores for: The **Electronic** Privacy Papers: **Documents** on the Battle for Priva of Surveillance by Bruce Schneier
www.comparebookprices.ca/book_detail/0471122971

5. Bill C-6: Personal Information Protection and **Electronic Documents**...

LS-344E BILL C-6: PERSONAL INFORMATION PROTECTION AND **ELECTRONIC DOCUMENTS** A
www.parl.gc.ca/36/2/parlbus/chambus/house/bills/summaries/c6-e.htm

6. Bill C-54: Personal Information Protection and **Electronic**...
LS-337E BILL C-54: PERSONAL INFORMATION PROTECTION AND **ELECTRONIC DOCUMENTS**
www.parl.gc.ca/36/1/parlbus/chambus/house/bills/summaries/c54-e.htm

7. Indexing Digital (**Electronic**) **Documents** -- It's Not an Option; Pay...
Electronic document management systems (EDMSs ... is not based on standardized **search** t
instance , a **search** that retrieves all **documents**...
fiat.gslis.utexas.edu/%7Escisco/inel.html

8. Key Topics - Protection of Personal Health Information
As of January 2002, the Personal Information Protection and **Electronic Documents** Act (PIPE
of Canada 2000, c. 5) applies to...
www.hc-sc.gc.ca/ohih-bsi/theme/priv/index_e.html

9. Translating Mathematical Markup into HTML
Translating Mathematical Markup for **Electronic Documents**. Keith Shafer Roger Thompson. A
Match_Parent restricts the context **search** to...
www.w3.org/pub/Conferences/WWW4/Papers/177/

10. Digital Discourses, On-Line Classes, **Electronic Documents**:
Digital Discourses, On-Line Classes, **Electronic Documents**: Developing New University Techn
Timothy W. Luke...
www.cddc.vt.edu/lol/html/Tim603.html

« Previous | Next »

Search for "**search electronic documents**" using: Google